Distributed Key-Value Stores: Performance and Scalability for Flash Media

Richard Elling
Key-Value (KV) Databases Overview

- Simple is important
  - Fast, easy to use
- Consistency is a differentiator
  - Strong consistency
    - Important for control planes
    - Examples: etcd, consul, zookeeper
  - Weak consistency (aka eventually consistent, NoSQL)
    - Popular for large-scale applications
    - Examples: dynamo, cosmos db
Kubernetes and KV

- Distributed system scheduler for containers
- How it works
  - Declare desired state in key-value database
  - Controllers (aka operators) watch the state and act to change current state to desired state
- End result
  - Billions of applications served reliably
  - Clusters up to 5,000 nodes each
Kubernetes Controllers/Operators

Observe
Watch cluster configuration database for interesting changes

Analyze
Current state = desired state?

Act
Do something to reconcile desired state

No
Observe
Watch cluster configuration database for interesting changes

etcd cluster config database

Analyze
Current state = desired state?

Act
Do something to reconcile desired state

No

I'm watching this
this changed
update this or that
## Etcd Cluster Configuration Database

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Etcd in Kubernetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency</td>
<td>Single-writer</td>
</tr>
<tr>
<td></td>
<td>Updates acknowledged by quorum of masters</td>
</tr>
<tr>
<td>Availability</td>
<td>Multiple-masters</td>
</tr>
<tr>
<td>Partition tolerance</td>
<td>Raft protocol ensures all masters are consistent</td>
</tr>
<tr>
<td>Performance</td>
<td>Writes committed to persistent storage log</td>
</tr>
<tr>
<td></td>
<td>Reads satisfied by any master</td>
</tr>
</tbody>
</table>
Etcd Persistent Storage

Client

put

node1

Raft

WAL

disk Δt

SSD

etcdserver

node1

Raft

WAL

disk Δt

SSD

etcdserver

node1

Raft

WAL

disk Δt

SSD

Client

put

node1

Raft

WAL

disk Δt

SSD
Kubernetes Practical Implementations

- Many nodes + fast SSD storage =
  - Many events to watch
  - When things break, many changes to process

- Can we improve etcd write workload scalability by adding faster storage?
etcd Benchmark

etcd operations throughput

- hdd: put and txn-put
- ssd: put and txn-put
- mem: put and txn-put

Flash Memory Summit 2019
Santa Clara, CA
Transaction Latency Analysis
Thank You

Richard.Elling@VikingEnterprise.com